George Kaceli

US Citizen | Windsor, ON | (519) 563-7858 | kaceli@uwindsor.ca | GitHub | LinkedIn | Portfolio

TECHNICAL SKILLS

Frameworks: ReactJS, Flask, Spring Boot, Django, Koa, Express, Ruby on Rails
Tools & Technologies: Docker, Kubernetes, Jenkins, Azure, GCP (Vertex AI, BigQuery), TCP/IP, REST API, Vim, PostgreSQL, GraphQL, Selenium, Locust, T-SQL, SSRS/SSIS
Programming Languages: Python, Java, C/C++, C#, Go, Ruby, JavaScript, Swift, R

EDUCATION

Master of Science Computer Science Artificial Intelligence Stream Co-op Sep 2024 - Present
University of Windsor, Windsor, ON

- Relevant Coursework: Deep Learning, Natural Language Processing, Software Engineering Topics

Bachelor of Computer Science, Honours Co-op with Software Engineering Specialization Sep 2019 - Jun 2024
University of Windsor, Windsor, ON

- GPA: 4.0

WORK EXPERIENCE

Agentic AI Engineer Co-op May 2026 - Sep 2026
Bell Canada

- Architected the grounding strategy for a text-to-SQL agent over contact-center analytics data, establishing deterministic metric resolution over vector retrieval and keeping the LLM out of the security boundary against prompt injection, cutting silently-incorrect SQL by 30% on an eval set
- Built a YAML metric glossary (typed expressions, enumerated categoricals, explicit grain rules) compiled from SME-authored Excel definitions and integrated it into a five-stage multi-agent pipeline built on Google Cloud Vertex AI (Agent Engine, ADK, Gemini, RAG Engine) over BigQuery, covering 40+ metrics across 3 domains
- Scaled the glossary hierarchically across 4 tables and extended it to a structurally distinct second domain (truck-roll data) with its own grain, metrics, and cross-domain joins, reducing new-domain onboarding time by 70%

Digital Transformation & Cloud Engineer Co-op Aug 2025 - Dec 2025
Municipality of Lakeshore, Lakeshore, ON

- Engineered and monitored secure cloud connectivity for 50+ municipal endpoints by supporting deployment and operation of the Cloudflare WARP client within a Zero Trust network architecture
- Designed and implemented advanced PowerShell-based network monitoring scripts to analyze UDP port activity (including UDP 443/QUIC), detect NAT timeout behavior, and assess tunnel stability for remote users
- Developed diagnostic tooling to measure MTU path constraints, packet fragmentation, and transport reliability across wired and wireless networks, reducing root-cause analysis time by 30%.
- Authored and maintained technical documentation in Confluence and tracked infrastructure issues and remediation tasks in Jira, contributing to standardized and auditable IT operations

Graduate Research & Teaching Assistant Jan 2023 - Present
University of Windsor, Windsor, ON

- Employed advanced techniques to fine-tune machine learning models such as BERT, LLAMA, and GPT, conducting rigorous model evaluation and time series analysis across various datasets to ensure robust application in real-world scenarios
- Established a feedback loop mechanism for fine-tuning LLM predictions, leading to a 20% improvement in question and answering accuracy and overall performance metrics gathered over 1 year
- Graded student assignments in the Design and Analysis of Algorithms ensuring concepts regarding correctness, time complexity and space were correctly applied
- Designated to provide guidance on algorithmic concepts, such as Dynamic Programming, Graph Algorithms and NP-Completeness

Software Developer Co-op Jan 2022 - Jan 2023
Ground Effects Ltd, Windsor, ON

- Constructed frontend features utilizing ReactJS and JavaScript, Node.js, Koa and GraphQL for the backend

- Created detailed paginated reports using SRRS and PowerBI enhancing data-driven decision-making for project managers
- Devised a server migration tool leveraging Python and SQL, automating the comparison of data during a server migration. Discovering and rectifying 100+ database anomalies, making database queries 15% faster
- Performed rigorous software testing on multiple applications, identifying over 15 critical bugs, increasing operational speed by 20%
- Built and optimized REST APIs using Kotlin and Spring, enabling seamless communication between systems

ACADEMIC PROJECTS

Enhancing Hierarchical RAG with Query-Aware Clustering (M.Sc. Thesis)

- Extending RAPTOR (Sarthi et al., ICLR 2024) hierarchical RAG by replacing its single Gaussian-mixture clustering step with five alternative clustering algorithms, including Leiden community detection, evaluated across three QA benchmarks (QASPER, QuALITY, NarrativeQA)
- Contributed a systematic clustering evaluation, structure-aware retrieval exploiting parent–child tree relationships, and an extractive-vs-abstractive summarization ablation, framed as indexing optimization within a Modular RAG framework
- Built self-contained per-benchmark evaluation harnesses (Token F1, multiple-choice accuracy, ROUGE-L) with tree caching and a fixed embedding model held constant across runs to keep the cross-method comparison valid

Transformer Based Automated Interlinear Glossing System

- Developed an interlinear glossing pipeline for low-resource languages with a character-level Transformer encoder, unsupervised morpheme segmentation, and a cross-attentive Transformer decoder
- Evaluated on SIGMORPHON 2023 Shared Task datasets achieving a micro-averaged Word-Level Glossing Accuracy of 72.51 % and Morpheme-Level Glossing Accuracy of 58.41 %, outperforming the shared-task baseline by over 27 % and 28 % respectively
- Demonstrated robust low-resource performance with greater than 80 % word-level accuracy on Natugu, Lezgi, and Tsez, and a 36.93 % word-level gain on Gitksan despite minimal training data

Condensed Variable Semantic Representation

- Constructed Variable Dependency Graphs (VDGs) from CodeSearchNet and fused pretrained VarCLR embeddings with graph structural features using a GCN encoder.
- Employed InfoNCE contrastive loss and a Deep Graph Infomax (DGI) objective to enforce both local and global mutual information objectives, achieving Spearman correlations of 0.47 (small split) and 0.45 (medium split) on IdBench.
- Introduced a fallback projection for out-of-graph variables, maintaining within 5 % of full-graph performance.

Fine-Tuning Large Language Models for Question Answering

- Fine-tuned pre-trained language models (BERT and Llama) on the SQuADv2.0 dataset, employing 4-bit quantization to reduce training time without sacrificing model performance
- Enhanced question and answering accuracy and efficiency through state-of-the-art fine-tuning techniques. achieving an Exact Match (EM) score of 82.13 % and an F1 score of 84.12 %

AWARDS & HONORS

- **Graduation with Great Distinction**, *University of Windsor: awarded for graduating with high academic standing, top 5% of class*
- **Dean's Renewable Entrance Scholarship**, *University of Windsor: merit-based scholarship awarded for sustained academic excellence*
- **Golden Key International Honour Society**, *Member: inducted for ranking in the top 15% of academic cohort*
- **Research & Academic Funding**, *University of Windsor: awarded competitive funding in support of graduate-level research in artificial intelligence and machine learning*